



Bayesian networks for knowledge discovery in a database from a program for genetic improvement of the Nelore breed

S.O. Rezende¹, C.A.J. da Rocha² & R.B. Lôbo³

*¹Institute of Mathematics and Computer Science,
University of São Paulo- São Carlos, SP, Brazil.*

*²University of Amazônia and Federal Center of Technological Education
of Pará, Belém, PA, Brazil.*

*³College of Medicine, University of São Paulo - Ribeirão Preto,
SP, Brazil.*

Abstract

The fact that companies are becoming more and more interested in acquiring new technologies for processing and storing data, as well as the view that information is their largest asset, has led to much research about the process of transforming this data into knowledge. In this context, the Knowledge Discovery in Databases process has emerged as a technology well suited to searching for knowledge that is embedded in the data. The work presented here involves the investigation of Bayesian concepts, techniques, methods and tools to aid a KDD process for a database from the Program for Genetic Improvement of the Nelore Breed (PMGRN). The main objective of the application of this process is to find coupling patterns for animals of this breed, considering some traits, such as weight, maternal ability, and scrotal circumference. Treatment of missing values is considered, as well as the incorporation of background knowledge into this process.



1 Introduction

The process of transforming data to knowledge has made use of mostly manual methods up to now to analyze and interpret the data, which makes the process of extracting patterns from databases quite expensive, slow and highly subjective, besides being almost impossible when working with large volumes of data. Interest in automating the analysis process for large volumes of data has fomented several projects in Knowledge Discovery in Databases (KDD) [1].

KDD consists of technology which statistical methods and tools, machine learning and databases to find a mathematical and/or logical description, eventually complex, of patterns and regularities in the data [1], [2]. This pattern search can be done using statistical methods and/or Artificial Intelligence techniques, especially those which manipulate uncertainty, and amply applied to data analysis with the objective of finding relationships and dependencies among the data. This article focuses on one of the pre-eminent ways of finding these relationships - Bayesian networks [3].

The main objective of this work is extract knowledge from the PMGRN database, especially anything that has to do with discovering coupling patterns for animals of this breed. The effects of incorporating background knowledge and the treatment of incomplete data is analyzed for this process.

This article is divided as follows: in Section 2, the context in which the KDD process is applied is introduced. In Section 3, the results obtained from the application of the KDD process to PMGRN database are presented. Finally, in Section 4, some conclusions are presented about this work.

2 Context of the Work

In 1997, the Computational Intelligence Laboratory (LABIC) at the Institute of Mathematical and Computer Sciences, through a partnership with the Genetics Department at the College of Medicine (DGFM) of the University of São Paulo, began work on a research project using technology for Knowledge Extraction from Databases. The objective was to find patterns and situations of interest by searching data from the PMGRN.

The Program for the Genetic Improvement of the Nelore Breed was started in June 1988, resulting from the union of cattle raisers searching for modern technologies that could be easily applied to the cattle farms, and researchers in the DGFM. Together they established goals to increase the productivity of the national beef herd.

According to the initial plan for using modern technologies, the Animal Model procedure is being used to evaluate bulls, heifers and young animals, thus obtaining Expected Progeny Differences (EPDs), for economic characteristics such as weight and scrotum circumference, standardized



for different ages.

By suggestion of the PMGRN experts one of specified objective is to find coupling patterns for animals considering some traits, such as weight, maternal ability and animal fertility.

3 Pattern Extraction Using the Data from PMGRN

Knowledge extraction from data can be understood as a process. This process starts with understanding the application domain, considering aspects such as the objectives of this application and the data sources from which knowledge is to be extracted. Next, a representative sample is taken from the database, pre-processed and submitted to the methods and tools of the Data Mining step, so as to find patterns/models (knowledge) in the data. Subsequently, this knowledge is evaluated as to its quality and/or usefulness so that, if the evaluation is positive, it can be used to support decision-making process. This article uses Bayesian network learning methods to extract these patterns (knowledge).

Bayesian Networks (BNs) can be seen as models that code the probabilistic relationships between the variables that represent a certain domain [5]. These models are made up of a qualitative structure, representing the dependencies between the nodes, and a quantitative one (conditional probability tables (CPTs) of these nodes), which evaluate these dependencies in probabilistic terms [3], [6]. Together, these components offer an efficient representation of the conjunct probability distribution (CPD) of the set of variables X for a domain [6]. The conjunct distribution is given by the following equation:

$$P(X) = \prod_{i=1}^n P(X_i | Pa_i) \quad (1)$$

in which Pa_i are the father-nodes of X_i . This representation makes it possible to substantially reduce the number of probabilities that are handled.

It is important to emphasize that the semantics of BNs facilitate, given the inherent causal representation of these networks, understanding and the decision-making process, for the users of these models. BNs can be built (or learned) using only the previous knowledge of the domain, the data or a combination of both. One of the main characteristics of this learning process is the possibility of making any kind of inference about the BNs, since all variables are treated indiscriminately.

3.1 Comprehension of the Application Domain

The PMGRN is currently made up of about 104 herds in 11 states in Brazil, with a total of about 173,000 animals. The animals that participate in the program are submitted to a periodic and very criterious control of their productive and reproductive characteristics. The information on these



traits are collected and sent (on floppy disks or over the Internet) by the cattle raiser, through a herd management program put together and supplied by the PMGRN Data Processing Center. All these data, along with the registration number of the animal, of its sire and its dam (for a total of 35 attributes) are stored in a database for analysis of interest to the cattle raisers and experts.

After getting an understanding of the PMGRN and the data that it contains, as well as the objectives to be accomplished by applying the KDD process, the program database was run through a process of selection and preparation, so as to meet the requirements of the Data Mining methods and/or the experts themselves.

3.2 Data Selection and Preparation

To accomplish the objectives proposed for the PMGRN data analysis, a new database had to be built, from the original one, taking into account the data of the animal and its parents.

This new database contains, in each register, all the attributes of the animal, its sire and its dam, for a total of 105 (35 x 3) attributes. A letter P was added to the attributes of the sire and a M was added to the attributes of the dam, to distinguish them from the attributes of their progeny. In this selection, only animals that had both sire and dam registered were considered (37,706 animals).

3.3 Data Mining

Before starting the pattern extraction process, and through talks with the expert, reductions were made in the attributes and instances of the resulting dataset. To make these reductions, a filter was established using the contemporary groups. The data were divided, by suggestion of the domain specialist, in contemporary groups, in which each dataset that resulted from this is formed by Nelore animals, born in the same year and submitted to the same handling conditions. Assuming that the objectives of the data analysis of the PMGRN are restricted to the EPD attribute values of the animals, for each dataset formed by a contemporary group, only the EPD attributes of these animals were considered. Thus, of the 105 attributes only 22 which referred to the EPDs were actually used.

Besides the division by contemporary group, a suggestion was made, by the domain experts, to subdivide this dataset by the traits growth, maternal ability and fertility. These traits were studied with the objective of finding patterns in the influence that the EPD values of the parents have on certain EPDs of their progeny. For each trait, only the more relevant EPDs were considered, besides the sex of the animal that matched each

Table 1: Dataset created with a dataset of animals born in 1994, considering traits suggested by the expert.

File	Records	Attributes	EPDs	Sex
Cres_94	6844	17	DDPP120 (EPDW120), DDPP365 (EPDW365) e DDPP550 (EPDW550)	Males and Females
Fert_94	3417	15	DDPE365 (EPDSC365)	Males
HMat_94	3427	15	DMPP120 (MAW120)	Females

trait. For example, for the dataset of the animals in the 1994 contemporary group, three sets of data were created for these traits, in which each one of these sets contained only animals with sex and EPD attributes related with a specific trait, as well as all the EPD attributes of the parents. The EPD attributes and the sex of the animals considered for the study of the traits, along with the datasets put together for this case study, can be visualized in Table 1. DDPP, DDPE, DMPP represent, respectively, EPDs for weight, scrotal circumference and maternal ability. After dividing them up by contemporary groups and traits, the data were placed in Bayesian Knowledge Discoverer.

The Bayesian Knowledge Discoverer (BKD) is capable of extracting Bayesian networks (structure + conditional probabilities) from data. It also implements a Goal-Oriented Propagation algorithm which can be used to support a decision-making process, quantifying it with probabilities. Inferences can be made in the generated Bayesian network, consulting the conditional probabilities, likelihood and variance, as well the marginal probabilities, given the occurrence of certain evidence in the states (values) of the variables of the network [4].

With the application of BKD knowledge was extracted in the form of dependencies between the EPDs of the parents and their progenies, making inferences to quantify these dependencies, besides studying the incorporation of background knowledge and treatment of incomplete data.

3.3.1 Bayesian Networks Generated by BKD

All the datasets generated for the contemporary groups that were analyzed (1990 to 1996) were supplied as input to BKD, and the BNs were generated to measure the correlations between the EPDs of the parents and of their progeny.

After generating the BNs, the BKD propagation algorithm was used to make inferences about the networks. In Figure 1, the BN generated with the HMat_94 (Maternal Ability) dataset is presented, along with the inferences made on it. As an example of inference, consider this situation: try to find the probability of an animal being part of an interval of higher values for DMPP120 [0.9, 3.6], given the evidence that DDPP120.P is in the set of values between [2.1, 4.3], and DMPP120.M has values between [-4.2, -0.5).

According with the Figure 1 the probability is 11%.

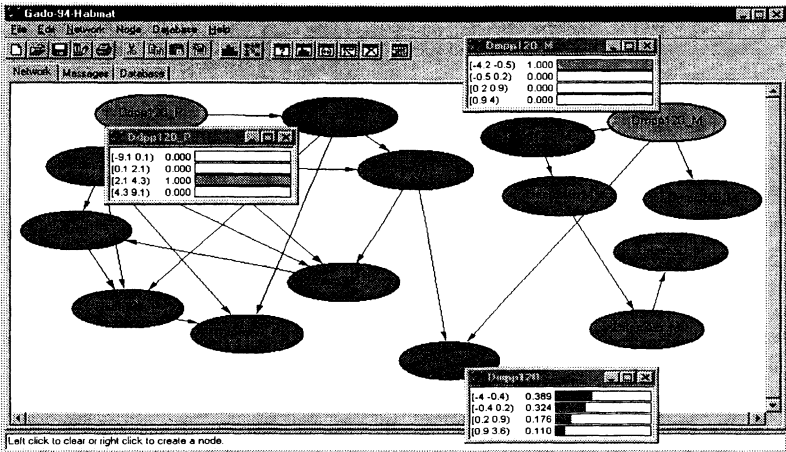


Figure 1: Inference made, with the propagation algorithm, on the Bayesian network generated by BKD using dataset Hmat.94.

Besides the BNs for the complete datasets, networks were also generated for the datasets with missing values and from the combination of background knowledge with other knowledge induced from the data.

3.3.2 Incorporation of Background Knowledge

Since BNs put together a model of the environment (of the domain of an application) and not just, as in many other representation schemes, a reasoning model, the background knowledge representation is particularly direct. Thus, the *a priori* knowledge can be represented by specifying elements which make up this model.

The BNs, along with the conditional probability tables, make it easy to combine background knowledge with the concepts extracted from the data, via techniques, studied and accepted in the literature, using Bayesian statistics [3]. The methods for learning BNs from data are used to refine the knowledge (structure and probabilities of the networks) provided *a priori*, according to the “concepts” embedded in the data.

In this work, background knowledge was incorporated considering *a priori* knowledge, given by the expert, about dependencies between the EPDs of the parents and their progenies, to generate BNs. The purpose of the incorporation of this background knowledge was to analyze and measure, through probabilities, some direct influences of certain EPD values of the parents related to their progeny, which were not observed by BKD, but rather assumed by the domain expert.

In Figure 2, an example is shown of including knowledge suggested *a*

priori. In this example, based on the BN generated with HMat_94 (Figure 1), an observation of the direct influence of DDPP120_P on DMPP120 of the progeny was suggested by the expert, even though it had not been observed by BKD.

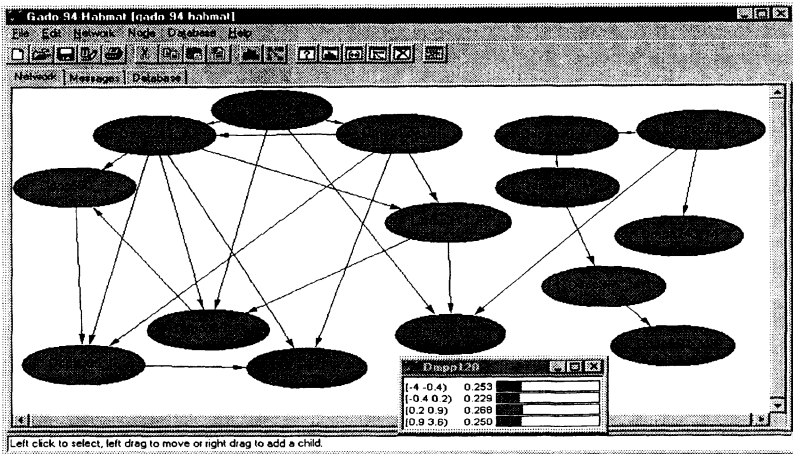


Figure 2: Bayesian Network generated using Hmat_94, considering the *a priori* knowledge that there is a dependency between DDPP120_P and DMPP120.

Notice that the CPT values of the DMPP120 node were altered, because of the change in the direct influence of the attributes. By incorporating the background knowledge, it was possible to explore many situations of interest, besides optimizing the time taken to build the and increasing the reliability of the results.

3.3.3 Treatment of Incomplete Data

In many real applications, the available data is incomplete, that is, it contains elements (attributes) with missing values. Kontraknen considers missing data an endemic problem when talking about databases [7].

These missing data can cause a series of problems during the process of extracting patterns and regularities from the data. To take care of this problem, several aspects should be considered, such as: in which step of the KDD process this will be taken care of mechanisms for generating missing values and the learning paradigm and type being used.

The effect of missing values on the representability of a sample can assume patterns depending on how the missing data are distributed in the sample. These patterns can be seen as mechanisms for generating missing values. The main ones are: missing at random (MAR), missing completely at random (MCAR) and not ignorable (NI) [4], [8]. A classification of the



generation mechanisms for incomplete data provides a way to understand the behavior of incomplete samples, but the problem of how to handle these missing values remains open, since ignoring the incomplete data mechanism does not imply that the missing values will be ignored.

It is important to notice that in the data of the database being used, missing values of the kind Don't Care and Informative Missing are not found, being treated as data which was not collected.

Of the different Bayesian methods available in the literature for treating absent values, the Bound and Collapse (BC) method was used in this work. This deterministic method estimates the parameters and the structure of a BN (with discrete variables) using a (possibly) incomplete database. The BC method is implemented in the software BKD, which can extract a BN from information available in a database, while also evaluating conditional probabilities using the extracted structure.

One of the centra points of this work was to evaluate the changes that occurred in the dependencies between the EPDs, given the existence of missing values in the dataset. 20 and 40% of the values of these EPDs were randomly removed from the dataset which was then submitted to BKD. The BN generated with dataset HMat.94, with 20% of missing values, has basically the same structure as the BN extracted from the complete dataset. The only new thing is some new dependencies between the nodes of the network. For example, the influence of DDPP120_M on DDPP550_M was added, and was not observed by the BN. It is interesting to note that the direct dependencies of the EPDs of the parents in relation to the EPDs of their progeny were not altered, or in other words, DMPP120 continued to be directly influenced by only EPDs DMPP240_P and DMPP120_M. Obviously, there were changes to the network parameters, as the frequency of occurrence of EPD values in the dataset given as input, was changed (20% of missing values).

3.3.4 Pre-Evaluation of the Results

The pre-evaluation of the results from the application of BKD was done using resources from BKD itself. The graphic environment of BKD was used to check for dependencies between the nodes of the generated BNs, analyzing the dependencies, as well as quantifying them using conditional and marginal probabilities, likelihood and variance. Besides all this, the propagation algorithm was used to make inferences, and to check for situations of interest, given the occurrence of certain evidence in the states (values) of the nodes (EPDs) of the network.

It is important to mention that the results obtained by applying the KDD process, using the PMGRN database, are in the final phase of evaluation by the experts.



4 Conclusions

This article describes a case study showing how background knowledge and treatment of incomplete data can be incorporated into a process of knowledge extraction from the PMGRN cattle database, using Bayesian networks.

The main contribution was the analysis of the PMGRN data to establish the correlations between the Expected Progeny Differences of the parents and their progeny. Knowledge extraction offer a real help to the prediction process, and consequently lower the costs of coupling and better progeny.

The interaction with the domain experts made it possible to better understand the PMGRN data. Also, because of the friendly user interface of the learning methods used, as well as the ease with which the models generated for the domain could be understood, we observed the great utility of these methods in helping the decision-making of those involved with the PMGRN, by including some of the preliminary results of this work in the Annual Summary of the Program [9].

In the experiments done by the authors it was possible to observe that the structure of the Bayesian networks changed very little over the years that were studied.

The results obtained indicate that the facility of incorporating background knowledge makes it possible to extract BNs closer to the "ideal" model, desired by the expert, for the PMGRN data.

Another important result is that direct dependencies between the EPDs of the parents and their progeny were maintained when 20% and 40% of the values of the original datasets were randomly removed. This indicates that missing values until 40% does not affect the BNs learning and consequently the prediction of new animals.

The experiment confirms that the EPDs of the parents are important for a good productive and/or reproductive performance of their progeny.

The preliminary results obtained by applying the KDD process encourage the use of this technology for determining the best coupling situations. This can be seen in the interest of the cattle raisers and experts in both hearing about and analyzing what was proposed with this work. Talks and workshops are currently being organized to promote the consolidation and use of the extracted knowledge.

Acknowledgements

The authors would like to thank cattle raisers of the PMGRN for providing us with their data. We are also grateful Prof. Maria Carolina Monard and students of two laboratories for their suggestions to this work. The authors acknowledge the support received from the research agencies CAPES, CNPq, FAPESP and FINEP.



References

- [1] Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P., The KDD Process for Extracting Useful Knowledge from Volumes of Data, *Communication of the ACM*, **39**, pp. 27-34, 1996.
- [2] Mannila, H., *Data Mining: Machine Learning, Statistics and Databases*, Department of Computer Science, University of Helsinki, 1997, URL: <http://www.cs.helsinki.fi/~mannila/>
- [3] Heckerman, D., Bayesian Networks for Data Mining. *Data Mining and Knowledge Discovery*, Kluwer Academic Publishers, **1**, pp. 79-119, 1997.
- [4] Ramoni, M. & Sebastiani, P., *Discovering Bayesian Networks in Incomplete Databases*, Knowledge Media Institute, The Open University, Technical Report, n. 46, 1997.
- [5] Russel, S. & Norvig, P., *Artificial Intelligence - A Modern Approach*, Prentice Hall, 1995.
- [6] Pearl, J., *Probabilistic Reasoning in Intelligent System*, Morgan Kaufmann Publishers, 1988.
- [7] Kontkanen, P., Myllymki, P., Silander, T. & Tirri, H., Comparing Stochastic Complexity Minimization Algorithms in Estimating Missing Data, *Proceedings Workshop on Uncertainty Processing - WUPES'97*, pp. 81-90, 1997.
- [8] Ghahramani, Z. & Michael, I. J., *Learning from Incomplete Data*, Artificial Intelligence Laboratory - Massachusetts Institute of Technology, 1994.
- [9] Rezende, S. O., Monard, M. C., Rocha, C. A. J, Oliveira, R. B. T., Nagai, W. & Caulkins, C. W. Extração de Conhecimento da Base de Dados do PMGRN, eds. Lôbo, R. B., Bezerra, L. A., Oliveira, H. N. & Los Reyes, A., *Avaliação Genética de Animais Jovens, Touros e Matrizes*, FMRP - USP / GEMAC (URL: <http://www-gen.fmrp.usp.br>), pp. 58-60, 1998.